Connectionism and compositionality: why Fodor and Pylyshyn were wrong

DAVID J. CHALMERS
Center for Research on Concepts and Cognition, Indiana University, Bloomington, Indiana 47405, USA

ABSTRACT *This paper offers a theoretical and an experimental perspective on the relationship between connectionist and classical (symbol-processing) models. A structural flaw in Fodor and Pylyshyn's argument against connectionism is pointed out: if, in fact, a part of their argument is valid, then it establishes a conclusion quite different from that which they intend, a conclusion which is demonstrably false. The source of this flaw is traced to an underestimation of the differences between localist and distributed representation. Fodor and McLaughlin have claimed that distributed representations cannot support systematic operations, or that if they can, then they will be mere implementations of traditional ideas. This paper presents experimental evidence against this conclusion: distributed representations can be used to support direct structure-sensitive operations, in a manner quite unlike the classical approach. Finally, it is argued that even if Fodor and Pylyshyn's argument that connectionist models of compositionality must be mere implementations were correct, this would still not be a serious argument against connectionism as a theory of mind.*

Introduction

In the last decade or so, connectionist models of cognition have proved successful at modeling diverse cognitive phenomena that have seemed resistant to the more traditional symbol-processing, or 'classical', cognitive models. Arriving at the height of the enthusiasm over these new models, Fodor and Pylyshyn's critique of connectionism (Fodor & Pylyshyn, 1988) threw a scare into the field, at least for a moment. A lively debate has ensued, with connectionists presenting a number of 'refutations', consisting in both theoretical arguments (e.g. Smolensky, 1987; Clark, 1989; van Gelder, 1990) and empirical counterexamples (Elman, 1990; Pollack, 1990; Smolensky, 1990). In turn, Fodor and McLaughlin (1990) have rejoined the fray for the classicists, with a reply to some of the arguments that connectionists have put forward.

In this paper I will offer a few observations on the issues. To begin, I will point out a structural flaw in Fodor and Pylyshyn's argument. Straightforward considerations about the structure of the argument will show that it cannot have succeeded in its intended purpose. Simple as these considerations are, they lead into deeper issues about just *why* the argument was wrong, and about the vital properties of connectionist models that were not taken into account. In particular, the role of *distributed representation* will be looked into. The ability of distributed representations to support structure-sensitive operations will be demonstrated with some experimental results, providing a counterexample to Fodor and McLaughlin's main argument. Finally, this will lead into the issue of the possible *implementation* of classical ideas by connectionist models, and expose the limited scope of some of Fodor and Pylyshyn's claims here.

A problem with Fodor and Pylyshyn's argument

Recall the major thrust of Fodor and Pylyshyn's argument: that connectionist models cannot support a compositional semantics. Or, more accurately, that they cannot support a compositional semantics unless they implement a classical architecture. Manifestations of compositional semantics are certainly ubiquitous in human thought, particularly in human language, through its *compositionality* (the meaning of "the girl loves John" is a function of the meaning of its constituent parts "the girl", "loves" and "John"), and its *systematicity* (the ability to think "John loves the girl" goes together with the ability to think "the girl loves John"). So if connectionism cannot handle compositional semantics, then that's a problem for connectionism.

The refutation of their argument can be stated in one sentence, then explained. *If their argument is correct as it is presented, then it implies that* no *connectionist network can support a compositional semantics; not even a connectionist implementation of a Turing machine, or of a Language of Thought.* This is a problem for Fodor and Pylyshyn, as it is well-known that connectionist networks can be used to implement Turing machines (or at least Turing machines with arbitrarily large but finite tape), and it is well-known that Turing machines can be used to support a compositional semantics [1]. Furthermore, the *human brain* is not unlike a connectionist network in many ways, and the human brain certainly supports a compositional semantics. So if their argument really establishes that *no* connectionist network can support a compositional semantics, then it establishes a false conclusion. So, applying the contrapositive of the italicized sentence above, Fodor and Pylyshyn's argument is not correct as it stands [2].

Of course, Fodor and Pylyshyn do not *intend* to imply such a conclusion. Indeed, they take great care to point out that the best future for connectionism lies in using it as an *implementational* strategy. Connectionist implementations of classical systems will certainly support a compositional semantics, albeit not in a particularly interesting way. In itself, this is a consistent position, and a sensible position for a classicist to take. However, Fodor and Pylyshyn's *position* is one thing. Their actual *argument* is another.

The substantive argument in Fodor and Pylyshyn's paper for the conclusion that connectionist models cannot support a compositional semantics takes up a relatively small portion of their paper (pp. 15–32). This starts with a simple localist connectionist network (that is, a network where each concept is represented by a single node, and each node represents a separate concept). They show that such a network cannot possess a compositional semantics, and argue that this applies equally to networks with distributed semantics (that is, a network with one concept being represented over many nodes). Therefore, the argument concludes, it is impossible for the semantics of a connectionist network to be compositional, whether this semantics is localist or distributed.

There is something strange about this conclusion. It is plainly false; it is universally recognised that *some* connectionist networks have compositional semantics: namely, connectionist implementations of classical architectures. So why are these not excluded from the argument? Going through the argument, the reader expects that at any point soon, there will be an *escape clause*—a clause showing why the argument as it stands does not apply to connectionist implementations of classical architectures. But this clause never appears; in fact, there is nothing close to such a thing. Fodor and Pylyshyn are left in the improbable position of having"proved" that even connectionist implementations of classical models have no compositional semantics. Faced with such a situation, we can only conclude that the argument is defective. Their supporters might argue that the flaw simply lies in the lack of an escape clause, which can easily be supplied; but no such escape clause is in evidence, and the onus lies with these people to provide it. In the meantime, we can conclude that the defect lies elsewhere: very likely, in the generalisation from localist to distributed semantics. More on this in a moment, after an analogy.

Say a mad scientist comes up to us with a "proof" that the Earth is the only inhabited planet in the universe. She runs through an impressive *a priori* argument, showing why it is impossible that the right kinds of biochemicals could be assembled in the right way, that the requisite organisational complexity could not arise, and so on. She concludes: life could not have arisen on any planet in this universe. But then, of course, it is an obvious fact that life arose on Earth. "That's OK," she answers, "that suits me fine. We knew that already. So what I've established is that life cannot have arisen anywhere but Earth". Now this will strike us as *ad hoc*, and as extremely poor logic. Her main argument never mentioned Earth; there was no *escape clause* showing just why the argument doesn't apply to Earth. To modify the conclusions of one's argument by considerations *external* to the argument is to admit that the argument is faulty. ("Mars is inhabited? OK, our argument demonstrates that life cannot have arisen anywhere but Earth or Mars".) If the argument can be fixed so that Earth is excluded from its force, very likely other planets will be excluded also. Analogously: if Fodor and Pylyshyn's argument can possibly be fixed up so that it excludes classical implementations from the scope of its conclusion, then the same fixes will probably exclude many other connectionist models too.

Before proceeding, I should note that some commentators have construed Fodor and Pylyshyn's argument somewhat differently from my construal here, interpreting is not as an argument specifically about connectionist networks, but as about the capacities of non-classical systems in general. On this construal, the argument goes: (1) systems without classical constituency relations between their representations cannot support a compositional semantics [3]; (2) only classical systems have classical constituency relations between their representations; therefore (3) no system can support a compositional semantics unless it implements a classical system. On this construal, an 'escape clause' is certainly built in. However, the first premise of this construal of the argument is blatantly question-begging, and is given little substantive support in the Fodor and Pylyshyn's paper [4]. Furthermore, this construal renders their detailed consideration of connectionist models (p. 15–32) entirely redundant. It therefore seems more natural to interpret Fodor and Pylyshyn's main argument as being concerned specifically with connectionist networks, so that the main burden of the argument is carried by the consideration and rejection (pp. 15–32) of the representational capacities of various kinds of connectionist model.

The refutation: a summary

All this has been a long-winded way of making the following simple argument:

(1) In Fodor and Pylyshyn's argument that no connectionist model can have compositional semantics, there is no escape clause excluding certain models (such as classical implementations) from the force of the conclusion (by observation).

(2) If their argument is correct as it stands, then it establishes that *no* connectionist model can have compositional semantics (from 1).

(3) But some connectionist models obviously *do* have compositional semantics; namely, connectionist implementations of classical models (by observation; accepted by all).

(4) Therefore, the argument is not correct as it stands (from 2 and 3).

Summing things up: let C denote the class of all possible connectionist networks, together with all possible associated semantics. Let FP denote the subset of C consisting of networks whose semantics are not compositional. Let L denote the subset of C consisting of networks with localist semantics. Let IMP denote the subset of C consisting of connectionist implementations of classical networks. The conclusion that Fodor and Pylyshyn *want* to establish is that $FP = C - IMP$.

In their argument, they first establish that $L \leq FP$. (Here "$\leq$" denotes set inclusion.) Let us grant them this, though some might argue. They then argue that it makes no difference whether the semantics are localist or distributed. Now, clearly the two possibilities of localist and distributed semantics exhaust the set C, so this argument, if correct, establishes that $FP = C$. But this is plainly false, as $IMP < C$ but it is not the case that $IMP < FP$.

We may conclude that *all* Fodor and Pylyshyn have established is that $L \leq FP \leq C - IMP$. The step in the argument that generalises to *all* distributed semantics is plainly defective. Although they would like to hold that it generalises to all distributed semantics *except* those used to implement classical models, the burden rests with them to show that this is the case. The conclusion established is a much weaker statement than $FP = C - IMP$. As things stand, it is just as likely that $FP = L$ as that $FP = C - IMP$, though no doubt the truth lies somewhere in the middle.

Localist and distributed representation

So far, we have demonstrated that Fodor and Pylyshyn's argument must be flawed. It remains to locate precisely the weak spot in the argument. To do this, we need to think about just *why* certain models—implementations and maybe others—slip through the argument's net. The answer is closely tied to the difference between localist and distributed representations.

Many connectionists have noted that the small localist network that Fodor and Pylyshyn used as their chief example is unrepresentative of the connectionist endeavor as a whole. When one asks what is the deepest philosophical commitment of the connectionist movement, the answer is surely this: the rejection of the atomic symbol as the bearer of meaning. Connectionists hold that atomic tokens do not carry enough information with them to be useful in modeling human cognition. Rather, distributed, subdivisible, malleable representations are the cornerstone of the connectionist endeavour. For this reason, many connectionists regard localist networks—whose basic representations are simple atomic units—as having more in common with traditional symbolic models than with distributed connectionist networks. The use of associative links means that these share some advantages with distributed models—soft constraint satisfaction, for instance—but they usually lack key features such as the ability to learn and to generalise automatically based on distance in a representational space, and they support a significantly less flexible and context-dependent style of processing.

The use of a localist network by Fodor and Pylyshyn, then, betrays a lack of understanding of the connectionist endeavor. F&P believe that nothing depends on the localist/distributed distinction; the connectionist, on the other hand, believes that everything depends on it. To F&P, a connectionist distributed representation is just a spread-out version of a single node (this comes out clearly in the footnote to p. 15). To the connectionist, a group of nodes functioning separately has functional properties far beyond those of an isolated unit. Small differences in the activity of a subset of nodes can make subtle or unsubtle differences to later processing, in a way that no single node can manage. A group of nodes carries has more internal structure than a single node, and as such to the connectionist is a far more likely candidate for semantic interpretation. And most importantly for our purposes here, whereas localist representations of a given concept are anchored to a single location within the functional organisation of the system, distributed representations of that concept can potentially float across different locations at different times, depending on context [5].

Before seeing precisely how distributed representations can overcome the limitations of localist representations, we should briefly how Fodor and Pylyshyn arrive at the conclusion that their argument applies equally to localist and distributed networks. The relevant material is brief. On the bottom of p. 15, we find:

> To simplify the argument, we assume a more 'localist' approach, in which each semantically interpreted node corresponds to a single Connectionist unit; but nothing relevant to this discussion is changed if these nodes actually consist of patterns over a cluster of units.

No substantive argument is provided here. And later (p. 19):

> To claim that a node is neurally distributed is presumably to claim that its states of activation correspond to patterns of neural activity—to aggregates of neural 'units'—rather than to activations of single neurons. The important point is that nodes that are distributed in this sense can perfectly well be syntactically and semantically atomic: Complex spatially-distributed implementation in no way implies constituent structure.

No one will begrudge Fodor and Pylyshyn this passage. As it stands, it is perfectly true. But it would only be interesting as argument if the last two sentences changed so that the "can" became a "must" and the "in no way implies" became "forbids". But it is precisely this that they cannot establish. We can conclude that their argument against distributed representation (and this is the extent of it) is weak. They go on to argue against connectionist models whose semantics are "distributed over microfeatures". But, as elsewhere, the kinds of semantics they consider bear little resemblance to those found anywhere in connectionism. This is the fundamental flaw in Fodor and Pylyshyn's argument: lack of imagination in considering the possible ways in which distributed representations can carry semantics. It is a different variety of distributed semantics that would be carried by a connectionist implementation of a Turing machine (and this, then, accounts for the logical flaw detailed above). And it is a different variety again of distributed semantics that can yield connectionist models of compositionality in important new ways.

To see just how distributed representations can escape the force of their argument, we must see how it is that connectionist implementations of Turing machines differ from localist connectionist networks in a way that enables compositional semantics. The relevant difference is straightforward. In a localist system, representation of a given concept is *anchored* to a specific location, so that it is impossible for such representations to move about the system,combining with other concepts as is necessary for compositional semantics. Given this anchoring, the only way a combination of these concepts can be represented is as an unstructured sum, by simply activating the relevant units for each concept. Therefore there will be no way to distinguish between different complex concepts that are conceptual combinations of the same constituent concepts. The proposition "The girl loves John", for instance, will be represented identically to "John loves the girl".

By contrast, if we look at classical systems—and in particular at connectionist implementations of classical systems—we find that concepts are not tied to any particular location, so they can be represented at different locations when entering into different conceptual combinations, enabling complex structures to be represented. Take a connectionist implementation of a Turing machine that in turn implements a deductive process in first-order logic. At the Turing-machine level, there will be no single tape-square that is devoted to the representation of a given predicate, for instance. A predicate will be represented by different tape-squares at different times, depending both on global context (e.g. the stage of the deduction) and local context (the concepts with which the predicate is being combined). This will be mirrored at the network-level: there will be no specific node to which the representation of the concept is tied.

It is precisely in virtue of this *movability* of representation that implementations of classical systems are able to support compositional semantics. And this movability of representation is precisely the relevant property that localist systems lack. Fodor and Pylyshyn's arguments on pp. 15–32 depend entirely on the assumption that connectionist representations are anchored; even when they consider distributed systems, they assume that a given concept will be tied to its own specific subset of nodes. Given this assumption, then compositional semantics will of course be impossible to achieve. However, distributed connectionism in general is subject to no such restriction. It is entirely possible to build distributed connectionist systems in which responsibility for the representation of a given concept can shift from one subset of nodes to another, depending on context, or even in which patterns of activation shift around subtly within a given set of nodes. In fact, some degree of movability of representation is automatic, given the well-known context-dependence of distributed representation.

For representation of a concept to be movable in this sense, we need not require that the concept be representable on entirely disjoint sets of units at different times. This will sometimes be the case (as for example in Smolensky's tensor-product system), but more commonly in connectionist systems, a concept will always be represented diffusely over the same set of units—e.g. a layer of hidden units—but in many different ways at different times, sharing that representational space with many other concepts. What is most important is that the concepts are not separately *anchored*—i.e. that there be no context-free configuration that always and only represents a given concept. Once we have allowed that concepts can be represented in different ways at different times, then we have opened the way for the representation of compositional structures, for example by letting the way a constituent concept is represented vary according to the concepts it is combined with, and according to its place in the overall structure that is represented [6].

Of course, movability of representation merely *allows* compositional semantics; it does not *guarantee* it. Standard first-generation connectionist systems do not support compositional semantics, as they do not exploit the movability of representation in the right way. But the possibility is there. Turning this possibility into actuality requires complex design work, but it turns out that this work has already been accomplished in some second-generation connectionist models.

In fact, if we examine some models put forward as *counterexamples* to Fodor and Pylyshyn's conclusion—those of Elman, Pollack and Smolensky—we find that all of these make use of the movability of distributed representation in an essential way. This is most obvious in Smolensky's tensor-product variable-binding system (1990), where a given concept will be represented on entirely different sets of nodes depending on the concept to which it is bound, with the result that complex hierarchical conceptual combinations can be represented. In Elman's "simple recurrent network" (1990), a given word as input will have different effects upon the representation, in the hidden layer, of a sentence of which it is part, depending on the word's position within the sentence, so that the representation of "Mary hit

John" is quite different from the representation of "John hit Mary". Pollack's RAAM (1990) has a similar property: when it represents complex trees, there is no fixed pattern for the representation of a given constituent concept; rather, representation differs depending on the concept's structural position within the tree. All these models exploit the non-anchored nature of distributed representation to escape Fodor and Pylyshyn's argument against connectionist models of compositionality.

Structure-sensitive operations on distributed representations

The classicist might now reply: "All this talk of distributed representations is all very well. Maybe you can *encode* compositional information into such a representation. But can you *use* it?" This point is initially plausible. If the structural information is present but cannot be processed, then it is useless. The classicist might hold that connectionist compositional structure might be buried too deeply, too implicitly, to be accessed in a useful way. Indeed, in a recent paper, Fodor and McLaughlin (1990) have argued that to support structure-sensitive processing, a compositional representation must have explicit tokens of the semantically constituent parts as its syntactic constituents (that is, the representation must be a *concatenation* of the representations of its semantic constituents). If this argument is correct, then connectionist representations that represent structure only in a distributed, implicit way will not have the causal power to support structure-sensitivity.

One obvious reply that the connectionist might make is that clearly *some* structure-sensitive operations can be supported by such representations: namely, the operation of extraction of the original constituents. Both Smolensky's and Pollack's models, for instance, include decoding processes that go from a compositional representation back to its parts. This reply, while correct as far as it goes, is not very interesting. If structure-sensitive processing must always proceed through an initial stage of decomposition into constituents, then what we are dealing with is essentially a connectionist implementation of a classical symbol processing. In such processing, distributed representation is used as a mere implementational technique.

Fortunately, this is not always the case. In fact, distributed representations of compositional structure can be operated on directly, without proceeding through an extraction stage. This offers the promise of a connectionist approach to compositionality that is in no sense an implementation of the classical notion.

I have performed a series of experiments demonstrating the possibility of effective structure-sensitive operations on distributed representations. I will only outline them briefly here; they are presented in more detail in (Chalmers, 1990). The experiments used a recursive auto-associative memory (RAAM; see Pollack 1990) to encode syntactically structured representations of sentences in distributed form. Following this, a back-propagation network learned to perform syntactic transformations directly from one encoded representation to another.

The sentences represented were all of similar syntactic form to "John loves Michael" (active) or "Michael is loved by John" (passive). Five different names/ verbs were used as fillers for each slot of subject, verb or object, giving 125 possible sentences of each type altogether. These sentences were assigned syntactic structure

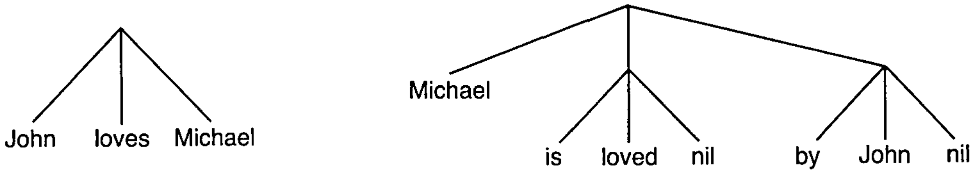

FIG. 1. Examples of sentences to be represented.

as shown in Fig. 1. A RAAM network was trained to encode 125 sentences of each kind into a distributed form. (Pollack, 1990 gives details of the RAAM architecture.) This is done by assigning each word a primitive localist representation (over 13 units), and then training a 39–13–39 backpropagation network (Fig. 2) to auto-associate on the three leaves descending from every internal "node" (in the trees in Fig. 1).

This gives us a 13-node distributed representation of the three leaves. Where necessary, this 13-node distributed representation is repropagated as part of the input to the 39–13–39 network, leading to higher-order structures being encoded. Eventually, we have a distributed representation of the entire tree. This process can be used, in principle, to encode any tree of valence 3 recursively.

The RAAM network learned to represent all 250 sentences satisfactorily, so that the distributed encodings of each sentence could be decoded back to the original sentence. These distributed representations were then used in modeling the process of syntactic transformation. In particular, the transformation of *passivisation* was modeled: that is, the passing from sentences like "John loves Michael" to sentences like "Michael is loved by John". (No commitment to any particular linguistic paradigm is being made here, and I make no claims for the psychological plausibility of this model. Syntactic transformations are used only as a clear example of the kind of structure-sensitive operation with which connectionist models are supposed to have difficulty.)

150 of the encoded distributed representations (75 active and the corresponding 75 passive sentences) were randomly selected for the training of the transformation network. This was a simple 13–13–13 backpropagation network (Fig. 3), which took a representation of an active sentence ("John loves Michael")

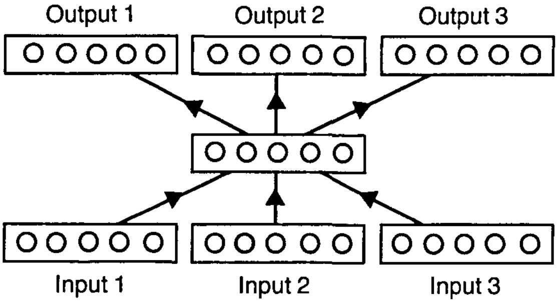

FIG. 2. The basis of the RAAM network.

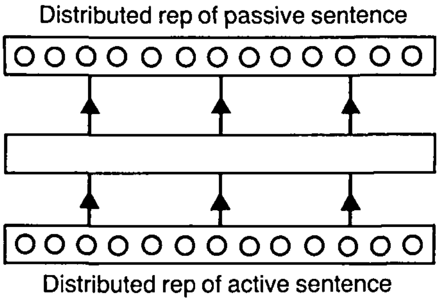
Distributed rep of passive sentence
Distributed rep of active sentence

FIG. 3. The transformation network.

as input, and was trained to produce a representation of the corresponding passivised sentence ("Michael is loved by John") as output.

Training proceeded satisfactorily. The interesting part was the test of generalisation, to see if the network was truly sensitive to the syntactic structure encoded in the distributed forms. The Transformation network was tested on the 100 remaining sentences from the original corpus. The 50 active sentences were encoded by the RAAM and fed to the transformation network, yielding a 13-node output pattern. This was fed to the RAAM network for decoding. In all 50 cases, the output pattern decoded to the correct passivised sentence. Thus, not only was the Transformation network able to be trained to optimal performance, but the generalisation rate on new sentences of the same form was 100%. The reverse transformation was also modeled (from passive to active). Performance was equally good, with a generalisation rate of 100%.

These results establish that it is possible for connectionist networks to model structure-sensitive operations directly upon distributed representations. This bears on the arguments at hand in two ways:

(1) It demonstrates that not only can compositional structure be *encoded* in distributed form, but that the structure implicitly present within the distributed form can be *used* directly for further processing. This provides a direct counterexample to Fodor and McLaughlin's claim. Despite the lack of explicit concatenative structure in the RAAM representations, they support structure-sensitive processing.
(2) It demonstrates the possibility of structure-sensitive operations in connectionist models that are in no sense implementations of classical algorithms. To see this, note that when a structure-sensitive operations is being performed upon a classical compositional representation, all processing *must* first proceed through a step of explicit decomposition, with particular tokens being explicitly extracted. In the connectionist model above, the transformation operation takes place without ever having to extract those constituent parts. Instead, the operation is direct and holistic.

It is important to note that this model is not intended as a model of *learning*. Although the model does in fact learn to produce systematic structure-sensitive processing, the means of learning is unrealistic, as people do not need to be exposed to more than half of the space of possible sentences in order to learn to perform transformations such as passivisation. Rather, learning is used here simply as a means to an end: the important result is the existence of the final network that performs structure-sensitive operations. However, learning and generalisation do play a subtle role in this demonstration. If the network had simply been wired by hand, or had been trained on all 125 possible sentences, then it would have been open to the accusation that its performance was not *systematic* at all; it might conceivably have been functioning as a mere unsystematic associator, like a look-up table that had memorised all 125 possibilities separately. The fact that the network generalised perfectly from a subset of 75 sentences to the full 125 shows that in fact this is not the case. If the network had simply learned the set of associations unsystematically, then we would expect generalisation to be no better than chance. The successful generalisation to novel sentences indicates that instead, its capacity to perform these associations was grounded in a systematic sensitivity to the sentences' structure.

Compositionality and nomologicality

Fodor and McLaughlin address the possibility of the kind of model I have presented only briefly, at the end of their article. This comes up in the context of discussing the ability of Smolensky's tensor-product representations to support structure-sensitive operations, given that this structure is not explicit but only "imaginary" in the structure of the representation:

> By way of rounding out the argument, we want to reply to a question raised by an anonymous *Cognition* reviewer, who asks: " . . . couldn't Smolensky easily build in mechanisms to accomplish the matrix algebra operations that would make the necessary vector explicit (or better yet, from his point of view, . . . mechanisms that are sensitive to the imaginary components without literally making them explicit in some string of units)?" But this misses the point of the problem that systematicity poses for connectionists, which is not to show that systematic cognitive capacities are *possible* given the assumptions of connectionist architecture, but to explain how systematicity could be *necessary*—how it could be a *law* that cognitive capacities are systematic—given those assumptions. (Fodor & McLaughlin, 1990, pp. 201–202)

The model that I have presented above corresponds to the possibility in parentheses: structure-sensitive processing without making the constituents explicit.

For the sake of argument, let us accept Fodor and McLaughlin's claim that compositionality is a law. The cash-value of this claim is presumably that under the conditions in which human minds develop, all developed minds exhibit compositionality of thought and language; we never find non-compositional minds, so presumably non-compositional minds are impossible, given standard developmental conditions. Fodor and McLaughlin's accusation is that although it is *possible* for models like this one to exhibit compositionality, it is also possible for them not to do so, and that this contradicts the nomological status of compositionality:

> No doubt it is possible for Smolensky to wire a network so that it supports a vector that represents aRb if and only if it supports a vector that represents bRa; and perhaps it is possible for him to do that without making the imaginary units explicit .. . The trouble is that, although the architecture permits this, it equally permits Smolensky to wire a network so that it supports a vector that represents aRb if and only if it supports a vector that represents zSq; or, for that matter, if and only if it supports a vector that represents the last of the Mohicans. (p. 202)

This argument seems to be a red herring. Given that the connectionist has demonstrated a class of systems that exhibit compositionality, systematicity, and so on, it is entirely irrelevant that there is another class of systems, wired up slightly differently, that do not. Those models are simply the *wrong* models, and do not fall in the relevant class under discussion.

It is true that for compositionality to be nomological, it must be preserved under certain counterfactual conditions. But counterfactuals about "if the network were wired up differently" are the wrong counterfactuals here. To see this, note that it is equally true of the human brain that if it were wired up slightly differently, then it might not support compositionality—but that affects the nomological status of human compositionality not a bit. Rather, the relevant counterfactuals concern different developmental conditions: that is, it is required that if the system developed under different environmental conditions, then it would still be compositional.

What a completed connectionist psychology will require, then, is (a) the exhibition of a network property P, such that any model with property P is compositional, systematic, and so on, and (b) an account of ontogenesis from some initial state, such that for any reasonable initial conditions and developmental environment, a model with property P will result [7]. So far, connectionist researchers have given (a) but not (b), for the simple reason that psychological development is still very poorly understood. So we can't yet evaluate the nomological status of compositionality in tensor-product and RAAM networks, as these do not come with associated accounts of development. But that is not to say that no such account can exist. For example, the connectionist might simply claim that property P is innate, and is not affected by development. More plausible might be an account in which property P lies in a basin of attraction along some developmental curve, under plausible environmental conditions. It will be a long time until we have satisfactory accounts of psychological development. The important point here is that Fodor and McLaughlin's argument has no bearing on the possibility of such an account.

The relationship between the approaches

A argument made frequently by Fodor and Pylyshyn is that connectionists have two choices: either (1) ignore the facts of compositionality and systematicity, and so have a defective theory of mind, or (2) accept compositionality and systematicity, in which case connectionism merely becomes a strategy for implementing classical models. The following passage is typical:

> if you need structure in mental representations anyway to account for the productivity and systematicity of minds, why not postulate mental processes that are structure sensitive to account for the coherence of mental processes? Why not be a Classicist, in short? (p. 67)

This argument is rather curious. It is not only that it contradicts the evidence, demonstrated above, that connectionism might model structure-sensitive processes in a non-classical way. There is also a deeply-embedded false assumption here: the assumption that *compositionality is all there is.*

To see the role that this assumption plays, shift the temporal position of the debate back a few decades. Let us imagine two traditional behaviorists, Fido and Pavlovian, who are distressed at the current turn of events. The revolutionary 'cognitivists' have recently appeared on the screen, and are doing their best to undermine the basic assumptions of decades of solid research in psychology. Our behaviorists have difficulty grasping the idea behind this movement. They express their bewilderment as follows: "Surely you all recognise that Classical Conditioning is a fact of human nature. The empirical evidence is overwhelming. But your cognitivist ideas do not take it sufficiently into account. There is no guarantee of stimulus-response association in your models as they stand. It seems to us that you have two choices: either (1) ignore the facts of Classical Conditioning, and therefore have a defective theory of mind, or (2) accept Conditioning and stimulus-response association, in which case cognitivism merely becomes a strategy for implementing the Behaviorist agenda" [8].

Presumably a cognitivist (such as Fodor or Pylyshyn) would quickly see the flaw in this argument. To be sure, conditioning is an empirical fact, and any complete theory must account for it. But it's certainly not the *only* fact, or even the most important fact, about the human mind. The cognitivists may pursue their own research agenda, making progress in many areas, and paying as much or as little attention to conditioning as they like. Eventually they will have to come up with some explanation of the phenomenon, and who knows, it may well end up looking something like the behaviorist story, *as far as conditioning is concerned.* But this doesn't mean that the cognitivist theory of *mind* looks much like the behaviorist theory overall, for the simple reason that *conditioning is only one part of the story.*

Similarly, compositionality is only one part of the story. Connectionists are free to pursue their own agenda, explaining various aspects of the mind as they see fit. Sooner or later, they will have to explain how compositionality fits into the picture. The story that connectionism tells about compositionality may prove quite similar to the classical picture, or it may prove different. But even if it proves similar, this diminishes the status of connectionism not at all. The fact that connectionism might implement classical theories of *compositionality* does not imply that connectionism would be implementing classical theories of *mind*. Compositionality is just one aspect of the mind, after all. (Aspects of cognition for which compositionality seems relatively unimportant include: perception, categorisation, motor control, memory, similarity judgments, association, attention,and many more. Even within those areas of cognition in which compositionality obviously plays an important role, such as language processing, it is still only part of the story.)

Behaviorism was very good at explaining conditioning, but it had a problem: it was *only* good at explaining conditioning. Fodor and Pylyshyn's classicism is good at explaining compositionality and compositional semantics, but it's not necessarily good at explaining much else. Both conditioning and compositionality are only small aspects of the mind; it seems to be an illusion of perspective that led to behaviorists and classicists putting so much respective emphasis on them.

Fodor and Pylyshyn's arguments establish that compositionality *exists*, but for their arguments above to succeed, they would need to establish a rather stronger claim: that compositionality is *everything*. Such a claim is obviously false, so connectionism can go on happily trying to explain those areas of the mind that it chooses to. If the connectionist story about compositionality ends up looking a little like the classical story, then well and good—that means that the classicists haven't been wasting their time completely all these years, and there may be room for a healthy amount of ecumenicism. In the meantime, preemptive relegation of either approach to a subsidiary role is probably a bad idea.

Acknowledgements

I would like to thank Stan Franklin, Bob French, Liane Gabora, Doug Hofstadter and Terry Horgan for helpful comments. An earlier version of this paper appeared in the *Proceedings of the 12th Annual Conference of the Cognitive Science Society*.

Notes

[1] Franklin and Garzon (1990) exhibit a connectionist implementation of a Turing machine.

[2] In an article I discovered after an early version of this paper was published, Horgan and Tienson (1987) present an argument closely related to this one, though in less detail.

[3] A representation A is a classical constituent of a representation B if (a) the representational content of A is a semantic constituent of the representational content of B, and (b) A itself is tokened as a syntactic constituent of B. Semantic constituency relations must be isomorphically mirrored by syntactic constituency relations. For example, in the English representational system, "girl" is a classical constituent of "John loves the girl". van Gelder (1990) calls this phenomenon "concatenative compositionality", as representation of a whole is a syntactic concatenation of representations of its parts.

[4] van Gelder (1990) gives a general argument to the effect that classical constituency relations among representations are not required for compositional semantics.

[5] The point that Fodor and Pylyshyn underestimate the power of distribution is by no means original. It was first made by Smolensky (1987).

[6] It is possible that even single-node representations could be movable in this sense, so long as they

were represented on different nodes at different times, depending on the context. Such a representational system would not be localist in the usual sense, as there would be no invariant node-to-concept correspondence, but it would not be strictly distributed either.

[7] One might argue that for a truly complete explanation, one must also give an explanation of how the initial state might be produced as a result of evolution by natural selection.

[8] We could bring a similar argument-parody to bear against Fodor and McLaughlin's nomologicality argument, above, as follows. "Sure", Fido and Pavlovian acknowledge, "your cognitivist models *permit* conditioning. But they do not *necessitate* it. One can easily design cognitivist systems in which the laws of conditioning do not hold. But conditioning is nomologically necessary, so cognitivist models are inadequate". The flaw in this argument should be clear.

References

CHALMERS, D.J. (1990) Syntactic transformations on distributed representations, *Connection Science*, 2, pp. 53–62.

CLARK, A. (1989) *Microcognition* (Cambridge, MA, MIT Press).

ELMAN, J.L. (1990) Structured representations and connectionist models, in: G. ALTMANN (Ed.) *Cognitive Models of Speech Processing: computational and psycholinguistic perspectives* (Cambridge, MA, MIT Press).

FODOR, J.A. & PYLYSHYN, Z. (1988) Connectionism and cognitive architecture: a critical analysis, *Cognition*, 28, pp. 3–71.

FODOR, J.A. & MCLAUGHLIN, B.P. (1990) Connectionism and the problem of systematicity: why Smolensky's solution doesn't work, *Cognition*, 35, pp. 183–204.

FRANKLIN, S. & GARZON, M. (1990) Neural computability, in: O. OMIDVAR (Ed.) *Progress in Neural Networks*, Vol. 1, Chapter 6 (Norwood, NJ, Ablex).

HORGAN, T.E. & TIENSON, J.L. (1987) Settling into a new paradigm, *Southern Journal of Philosophy*, 26, Supplement, pp. 97–114.

POLLACK, J.B. (1990) Recursive distributed representations, *Artificial Intelligence*, 46, pp. 77–105.

SMOLENSKY, P. (1987) The constituent structure of connectionist mental states: a reply to Fodor and Pylyshyn, *Southern Journal of Philosophy*, 26, pp. 137–163.

SMOLENSKY, P. (1990) Tensor product variable binding and the representation of symbolic structures in connectionist systems, *Artificial Intelligence*, 46, pp. 159–216.

VAN GELDER, T. (1990) Compositionality: a connectionist variation on a classical theme, *Cognitive Science*, 14.